Multiple Neural Network Classification Scheme for Detection of Colonic Polyps in CT Colonography Data Sets[1]

Anna K. Jerebko, PhD, James D. Malley, PhD, Marek Franaszek, PhD, Ronald M. Summers, MD, PhD

Rationale and Objectives. A new classification system for colonic polyp detection, designed to increase sensitivity and reduce the number of false-positive findings with computed tomographic colonography, was developed and tested in this study.

Materials and Methods. The system involves classification by a committee of neural networks (NNs), each using largely distinct subsets of features selected from a general set. Back-propagation NNs trained with the Levenberg-Marquardt algorithm were used as primary classifiers (committee members). The set of features included region density, Gaussian and mean curvature and sphericity, lesion size, colon wall thickness, and the means and standard deviations of all of these values. Subsets of variables were initially selected because of their effectiveness according to training and test sample misclassification rates. The final decision for each case is based on the majority vote across the networks and reflects the weighted votes of all networks. The authors also introduce a smoothed cross-validation method designed to improve estimation of the true misclassification rates by reducing bias and variance.

Results. This committee method reduced the false-positive rate by 36%, a clinically meaningful reduction, and improved sensitivity by an average of 6.9% compared with decisions made by any single NN. The overall sensitivity and specificity were 82.9% and 95.3%, respectively, when sensitivity was estimated by means of smoothed cross-validation.

Conclusion. The proposed method of using multiple classifiers and majority voting is recommended for classification tasks with large sets of input features, particularly when selected feature subsets may not be equally effective and do not provide satisfactory true- and false-positive rates. This approach reduces variance in estimates of misclassification rates.

Key Words. Colon neoplasms, CT; colon neoplasms, diagnosis; computers, diagnostic aid; computers, neural network.

Computed tomographic (CT) colonography as an alternative colon cancer screening technique has progressed rapidly over the past 6 years (1). Colon cancer remains a serious risk, affecting approximately 6% of Americans during their lifetime. Total colon evaluation with widely accepted methods such as barium enema studies and conventional colonoscopy can identify most polyps before they progress to cancer. Nevertheless, a large portion of the population older than 50 years do not undergo screening because they fear the discomfort of these screening tests. The relatively high cost of the examination and difficulties in reaching the most distant parts of the colon are also problems with the standard test, conventional colonoscopy. For these reasons, there is a need to develop accurate, less invasive, relatively inexpensive, and potentially more powerful screening methods, such as CT colonography.

Acad Radiol 2003; 10:154–160

[1] From the Department of Diagnostic Radiology, Warren Grant Magnuson Clinical Center, National Institutes of Health, Bldg 10, Rm 1C660, 10 Center Dr, MSC 1182, Bethesda, MD 20892-1182 (A.K.J., M.F., R.M.S.); and Center for Information Technology, National Institutes of Health, Bethesda, Md (J.D.M.). Received September 5, 2002; revision requested October 28; revision received and accepted November 7. Supported by the intramural research programs of the Diagnostic Radiology Department, Warren G. Magnuson Clinical Center. **Address correspondence to** R.M.S.

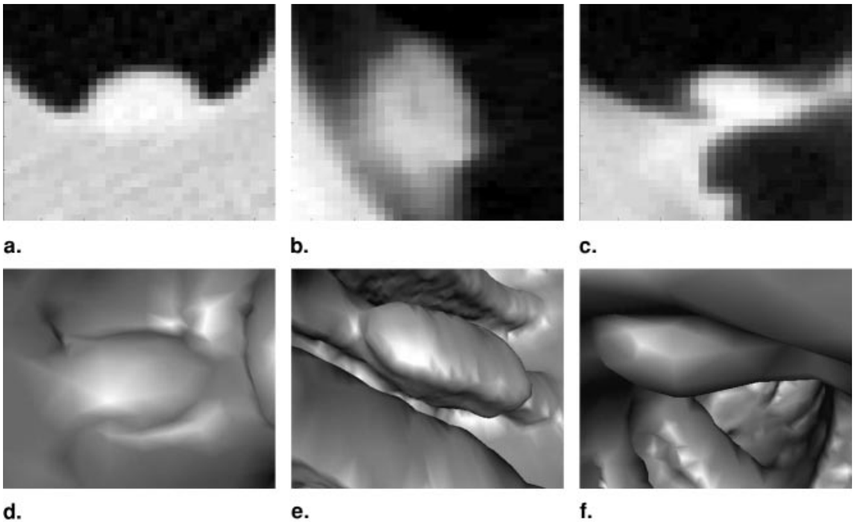


Figure 1. **(a)** Sessile polyp. **(b)** Polyp on a fold. **(c)** Pedunculated polyp lying on its side. **(d–f)** Three-dimensional images of the same polyps.

Computer-aided detection may improve the accuracy and reproducibility of CT colonography (1). Such detection for CT colonography is at an early stage of development. One important goal of computer-aided detection is to optimize the classification scheme employed. The classification method for colonic polyps discussed in this article combines two techniques: a shape-based primary classifier and a higher-level classifier based on an aggregate of back-propagation neural networks (NNs) (2). Three-dimensional shape plays a key role in human visual perception of objects in general and in colonic polyp detection specifically. The shape-based classifier for primary lesion detection eliminates up to 97% of the colonic surface from consideration (3). The remaining 3% represents on average 65 candidate lesions per colon, which include a mixture of real polyps (nearly 100% of the true polyps being included) and false-positive detections. These candidate polyps are thus suitable for further processing with a higher-level classification scheme.

While shape criteria are a source of primary features (variables used for prediction) known to be clinically useful, other features considered in the diagnosis of polyps are more difficult to quantify. In this situation it is helpful to base decisions on as many features as possible. In our study, we tested the effectiveness of an aggregate committee, or voting bloc, of several NNs for classifying polyps. Each committee member used a different set (collection) of four features selected from the general list of up to 23 features. A decision for a test polyp was based on the majority vote of the committee members. We hoped to achieve greater parallelism and higher sensitivity and specificity rates with this method than had been achieved previously.

MATERIALS AND METHODS

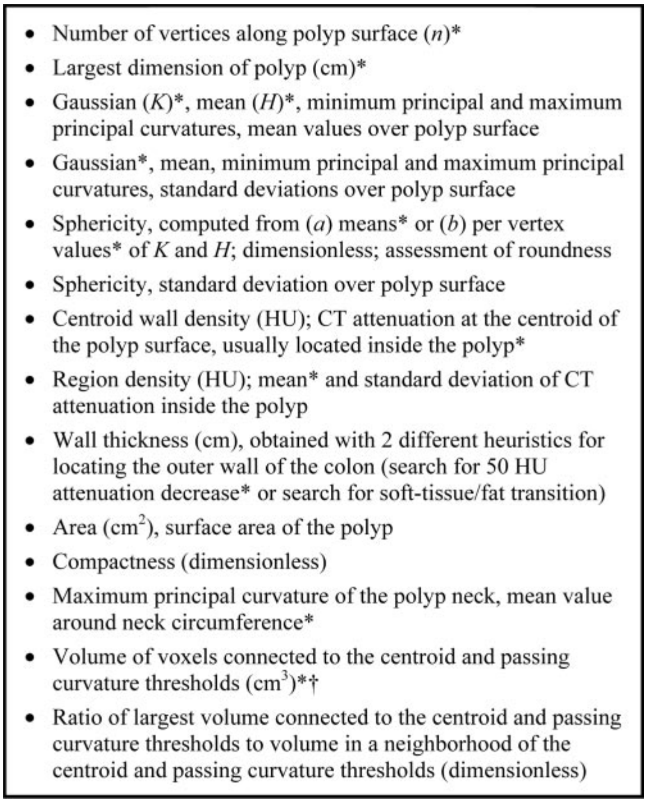
Shape Classification

The first step of the classification algorithm is meant to eliminate most of the colonic surface, which is unlikely to contain true polyps. It involves analyzing the geometric shape features of the colonic surface, such as curvature and sphericity. Although colonic polyps may vary in size and shape (eg, pedunculated, hyperplastic, and sessile [4]), most appear as bumps on the computer-rendered image of the colon surface. Examples of the different polyp shapes are shown in Figure 1.

The algorithm first selects lesions that have elliptical curvature and sphericity above certain thresholds. Then other important characteristics are calculated, such as size, region density, and wall thickness, which improve specificity without serious loss in sensitivity.

Aggregate of Back-Propagation NNs

The next step in the classifier involves various texture, density, and geometric parameters of the lesions, colon surface, and wall, as well as their means and standard deviations. The complete collection of additional variables includes 23 features, some of which had proven useful for classification in earlier work (3) and some of which we added for this study (Fig 2). On the one hand, the inclusion of more features allows more precise discrimination between true- and false-positive detections. On the other,



- Number of vertices along polyp surface (*n*)*
- Largest dimension of polyp (cm)*
- Gaussian (*K*)*, mean (*H*)*, minimum principal and maximum principal curvatures, mean values over polyp surface
- Gaussian*, mean, minimum principal and maximum principal curvatures, standard deviations over polyp surface
- Sphericity, computed from (*a*) means* or (*b*) per vertex values* of *K* and *H*; dimensionless; assessment of roundness
- Sphericity, standard deviation over polyp surface
- Centroid wall density (HU); CT attenuation at the centroid of the polyp surface, usually located inside the polyp*
- Region density (HU); mean* and standard deviation of CT attenuation inside the polyp
- Wall thickness (cm), obtained with 2 different heuristics for locating the outer wall of the colon (search for 50 HU attenuation decrease* or search for soft-tissue/fat transition)
- Area (cm^2), surface area of the polyp
- Compactness (dimensionless)
- Maximum principal curvature of the polyp neck, mean value around neck circumference*
- Volume of voxels connected to the centroid and passing curvature thresholds (cm^3)*†
- Ratio of largest volume connected to the centroid and passing curvature thresholds to volume in a neighborhood of the centroid and passing curvature thresholds (dimensionless)

Figure 2. Features used for computer-aided detection with CT colonography. The 12 features used in the final NN committee are indicated by an asterisk (*). The feature that had two input nodes in the large single NN, giving it added weight, is indicated by a dagger (†).

using too many features in any single classifier (NN, classification tree, or other scheme) unacceptably increases the complexity of the model. With NNs, the number of hidden neurons (2) for an NN classifier corresponds to the dimensionality of the feature space. Keeping this dimensionality small effectively controls the model's complexity and increases the accuracy of parameter estimation (2,5). Here we suggest breaking the set of features into subsets and using a combination of several simple classifiers, each processing a small number of input features, so that each classifier works inside a reduced feature space. This approach combines the advantages of using the large number of features and keeping the feature space small for each NN in the committee. The entire polyp detection scheme is depicted in Figure 3.

A committee approach to classification is known to produce generally improved results, provided that error rates are less than 50% for each member of the committee. A simple example makes this point (6). Consider a collection of 21 classifiers, each with an error rate of 0.3 (or a sensitivity of 70%). With the standard binomial probability table, one can verify that if the classifiers are statistically independent, the probability is 0.026 that 11 or more of them will make a mistaken classification (sensitivity, 97.4%). Therefore, with even moderately efficient classifiers that are at least approximately independent, the classification error rate can drop dramatically. Choosing relatively distinct, disjoint feature sets from which to generate individual classifiers is one way to produce relatively independent classifiers and thereby reduce the committee error rate. Such selection is an ad hoc process. In this study, the process was guided in part by clinical insight and our experience with pairwise measures of independence among features. It might be desirable to make these selections with more refined methods, such as so-called genetic or evolutionary selection algorithms (7).

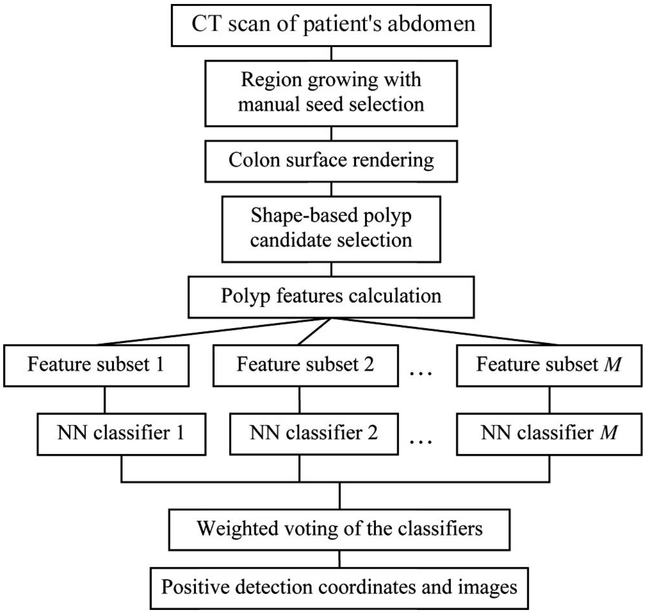

CT scan of patient's abdomen
Region growing with manual seed selection
Colon surface rendering
Shape-based polyp candidate selection
Polyp features calculation
Feature subset 1
Feature subset 2
…
Feature subset *M*
NN classifier 1
NN classifier 2
…
NN classifier *M*
Weighted voting of the classifiers
Positive detection coordinates and images

Figure 3. Polyp detection scheme.

The feature space *F* is divided into *M* subsets, each containing *N* features:

$$F \Rightarrow \begin{bmatrix} f_{11} \\ f_{12} \\ \cdots \\ f_{1N} \end{bmatrix}, \begin{bmatrix} f_{21} \\ f_{22} \\ \cdots \\ f_{2N} \end{bmatrix} \cdots \begin{bmatrix} f_{M1} \\ f_{M2} \\ \cdots \\ f_{MN} \end{bmatrix}. \tag{1}$$

These subsets are the input for the component NN classifiers. The features may be repeated in the different subsets, so many fairly disjoint combinations of them may be used. The individual NN classifiers are multilayer per-

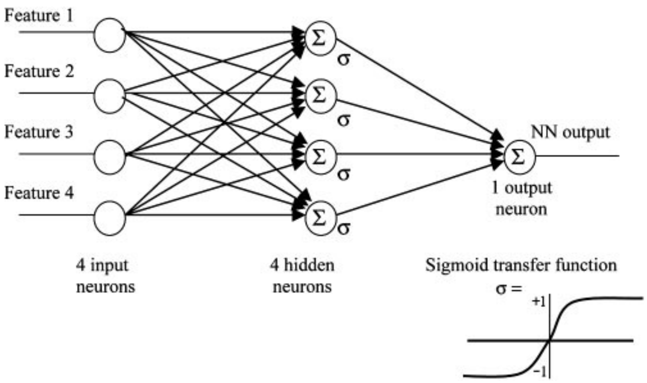


Figure 4. Diagram of the NN.

ceptrons trained with a back-propagation algorithm. Each NN consists of an input layer with N input neurons, a hidden layer with 2^q hidden neurons (where q is an integer), and an output neuron. The hidden neurons are connected to all input neurons and an output neuron, as depicted in Figure 4. The polyp data set contains K samples. M feature vectors of size N are calculated for each polyp. The output O_{jk} of the jth hidden neuron when kth sample X is presented is calculated as follows:

$$O_{jk} = \sigma\left(\sum_{i=1}^{N} w_{ij}X_{ik} + b_j\right), \tag{2}$$

where w_{ij} = the weight of connection, b_j = the bias of connection, $j = 1, \ldots, 2^q$, and σ = a sigmoid transfer function. In our experiments, the best results and fastest convergence were obtained for $q = 2$.

The output of the network g_k when the kth sample is presented is calculated as follows:

$$g_k = \sum_{j=1}^{2^q} w_j^{\text{out}} \cdot O_{jk} + b_{\text{out}}\,, \tag{3}$$

where w_j^{out}, b_{out} are the weights and the bias of the output neuron connections to all neurons of the hidden layer.

The goal of the training process is to have the NN output equal to -1 when the sample presented to the input layer represents a false-positive detection and equal to $+1$ when a true-positive detection is presented. Let V denote the weights and the biases of the NN: $V = V\,(w_{ij}, b_{j,}\, w_j^{\text{out}}, b_{\text{out}})$. Then the risk functional is the mean-squared error (MSE) function:

$$R = \frac{\sum_{k=1}^{K} [g(X_k, V) - y_k]^2}{K}, \tag{4}$$

where $X_k(f)$, y_k = the kth sample, $y_k = 1$ if $X_k(f)$ corresponds to a true-positive detection, and $y_k = -1$ if $X_k(f)$ corresponds to a false-positive detection. We used the Nguyen-Widrow method (8) to initialize V. An advanced nonlinear Levenberg-Marquardt optimization algorithm (9) was used to train the weights and biases V. We used MatLab computing software (The MathWorks, Natick, Mass) for NN simulation and training. Training was iterative and stopped when the desired MSE was reached.

Voting System for NN Committee

The classification model containing trained NNs uses weighted voting of all NNs in such a way that those with the weakest performance contribute least to the final decision. The weight attached to each NN is based on the numbers of false-positive and false-negative responses found when that trained NN is applied to both training and test sets. The weight (P) is calculated as follows:

$$P = \frac{1}{1 + \frac{k_1 \cdot \text{Fn}_{\text{tr}}}{N_1} + \frac{k_2 \cdot \text{Fn}_{\text{test}}}{N_2} + \frac{k_3 \cdot \text{Fp}_{\text{tr}}}{N_3} + \frac{k_4 \cdot \text{Fp}_{\text{test}}}{N_4}}, \tag{5}$$

where k_i, $i = 1, \ldots, 4$, are coefficients adjusted according to clinical needs. For better sensitivity, k_1 and k_2 should be higher than k_3 and k_4, and if specificity is more important for a particular study, then k_3 and k_4 should be higher. N_1 and N_2 are the numbers of polyps in the training and test sets, respectively, and N_3 and N_4 are the corresponding numbers of nonpolyps. Fn_{tr} and Fn_{test} are the numbers of false-negative responses found when the trained NN is applied to training and test sets, respectively, and Fp_{tr} and Fp_{test} are the corresponding numbers of false-positive responses.

In the results we describe, we used a simple threshold scheme: If a given NN had P greater than the threshold, then it was included in the committee; if its P was at or below the threshold, it was dropped. This procedure reduced the number of features finally used in the winning subsets; that is, the starting set of 23 features was trimmed to a smaller set of 12 features that was used in the final committee of NNs (Fig 2). The weights of all included NNs were then set to $+1$, and simple majority vote then followed. We do not assert that our weighting scheme is statistically optimal, and many other choices are possible.

To get the best estimates of the NNs' weights (P) and to assess feature selection, we used a bootstrap smoothing of the basic leave-one-out cross-validation method (10). This approach gives error estimates with relatively low bias and reduced variance, compared with a nearly unbiased but highly variable traditional cross-validation (either leave one out or leave $k > 1$ out).

The smoothed leave-one-out method was implemented as follows. A single true polyp was chosen from the list of true polyps K in the training set. From the remaining true polyps $K - 1$, a set of cases $K - 1$ was drawn, where drawing was done with replacement; this formed a single bootstrap draw. From the nonpolyps, a drawing was done with replacement to obtain L cases. The bootstrap set of $K - 1$ drawn true polyps and L nonpolyps was a new training set, which we used to train the eight NNs. The committee of these eight networks then voted on the initially selected polyp, and we recorded the result. This process was repeated $M = 10$ times. Next, the chosen true polyp was replaced and another true polyp was selected. The bootstrap drawing was repeated, and the votes were collected. In this way each of $K = 21$ polyps in our data set was used $M = 10$ times to test the NNs trained on the 10 bootstrap replications of the remaining 20-polyp training sets. The final error rate estimates and the weight for each NN (value of the P function above and the threshold value of P) were obtained by averaging over all 210 tests. The NN committee classification decision D was calculated as follows:

$$D = \sum_{i=1}^{M} g_i P_i, \tag{6}$$

where g_i is the output of the ith NN, P_i is the ith NN weight, and M is the number of NNs in the committee.

Experimental Data

The polyp database used for training and test purposes in our experiment was obtained from 80 studies that included supine and prone CT colonographic images of 40 patients (11). Informed consent was obtained from the patients. CT imaging was performed with Lightspeed scanners (GE Medical Systems, Milwaukee, Wis) at the following parameters: 120 kVp, 50 mAs (mean), field of view to fit (38–46 cm), 5-mm collimation, high-quality mode, and 3-mm reconstruction interval (2-mm overlap). We performed colonoscopies on the same patients to verify the results of CT scans and obtain the coordinates of true-positive findings. Twenty-one polyps (0.5–2.5 cm) and 4,996 polyplike sites were selected with our endoscopic research software (3) according to the shape criteria.

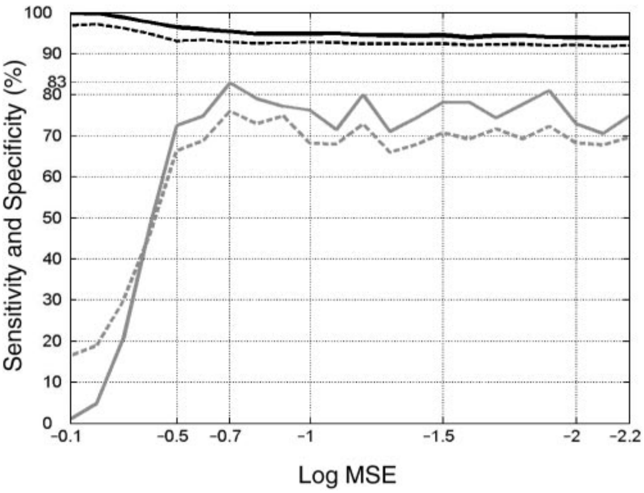


Figure 5. Specificity and sensitivity of the cross-validated NN committee (black and gray solid lines, respectively) compared with the specificity and sensitivity of the average single NN (black and gray dotted lines, respectively).

RESULTS

We used the smoothed leave-one-out validation technique described earlier in this article to analyze the sensitivity, specificity, and variability of an NN committee and a single NN, at each of 22 different MSE levels, with log MSE from −0.1 to −2.2. Log MSE is used as a stopping criterion in NN optimization. We used a range of log MSE values to guide in the training of NNs with optimal sensitivity and specificity. To reduce the computational cost of the analysis, we devised a committee that consisted of eight NNs with the highest weights (according to the threshold method described previously). The final committee used only 12 input features, combined in eight different sets of four features each.

Analysis has shown that at high MSE levels, when log MSE is greater than −0.4 but less than −0.1, a single average NN and an NN committee classifies the majority of samples as nonpolyps, giving high specificity but very low sensitivity. As can be seen from Figure 5, the com-

Performance of the NN Committee and the Average Single NN at Log MSE = −0.7

	Single NN with Four Input Nodes										
	NN 1	NN 2	NN 3	NN 4	NN 5	NN 6	NN 7	NN 8	Average	NN Committee	Single NN, 13 Input Nodes
No. of false-positive results per study	4.1	4.4	4.6	4.1	6.1	4.3	5.0	3.8	4.6	2.9	4.5
Specificity (%)	93.5	93.0	92.7	93.4	90.2	93.2	92.0	93.9	92.7	95.3	92.7
Specificity internal STD (%)	0.7	0.7	0.4	0.4	0.8	0.6	0.6	0.5	0.6	0.2	0.7
Sensitivity (%)	81.4	74.8	75.2	80.5	59.1	78.6	80.5	77.6	76.0	82.9	57.6
Sensitivity internal STD (%)	6.1	8.4	8.6	6.9	10.1	7.5	6.5	5.0	7.4	6.0	8.8

Note.—For each bootstrap sample (draw with replacement), a single NN makes a classification on a leave-one-out case. The average error over the 21 true polyps is found. This is done for each of 10 bootstrap draws and the standard deviation (STD) of these 10 error rates is what we call the internal STD. This is also true for the committee and the large single NN having 13 input nodes.

mittee specificity is always higher than the average specificity of any single NN, at each MSE. The committee sensitivity is higher than that of an average single NN when log MSE is at most −0.3.

Sensitivity reaches its maximum for single NNs and an NN committee at a log MSE level of −0.7. At this point the committee sensitivity estimated with the smoothed leave-one-out method is 82.9%, while the average sensitivity of the eight single NNs is 76.0% (Table). The specificity of the NN committee at this same log MSE level is 2.6% higher than that of the average single NN, yielding a 36% improvement in the number of false-positive results per study (2.9 vs 4.6). We also analyzed the performance of a single large NN with 12 input features (all the features used in eight smaller NNs composing the committee [Fig 2]). The sensitivity, specificity, and false-positive rate (per study) for this single NN are given in the far-right-hand column of the Table. The sensitivity for this NN is less impressive than the average for the smaller NNs that use only four features each; all performance indicators for this NN are much less impressive than performance indicators for the NN committee. In particular, the standard deviation for specificity is 0.2% for the committee 0.6% on average for the single networks, and 0.7% for the big NN. Similarly, the standard deviation for sensitivity was 6.0% for the committee, 7.4% on average for the single networks, and 8.8% for the big NN.

We found that both the NN committee and the single NN classification models become overtrained with log MSE levels lower than −0.7; sensitivity rates decreased rapidly. This is the well-known problem of overfitting, in which the classification model works nearly perfectly on the training set but gives poor results on a test set.

DISCUSSION

In this study, we found that a committee of NNs improved sensitivity and specificity by 6.9% and 2.6%, respectively, compared with any single NN. The result was a 29% reduction in false-negative detections and a 36% reduction in false-positive detections.

To reduce the training time, we used a smaller training set of 100 nonpolyps and 21 true polyps. This was necessary because the intensive validation scheme that provided the improved estimates of error rates was itself computationally expensive: In our study, eight NNs (each based on four of 12 features, with some features being used in more than one NN) were trained 210 times, each to 22 different MSE levels, for validation purposes and then used to select a final committee. For clinical applications, however, each NN in the committee has to be trained just once. Since this takes less than a minute on a Pentium III desktop computer, an arbitrarily large training set can be used to reach the required sensitivity and specificity. In other words, only the validation process used here was computationally expensive; application of the committee scheme itself was not expensive.

Our results lead us to recommend the NN committee for classification of polyps. We observed that the smoothed bootstrap validation approach demonstrated average improvements of 6.9% in sensitivity and 2.6% in specificity. The problem of evaluating statistical differences between competing classification rules has been discussed by Dietterich (12), who also suggested an alternative—five splits of the data into two subsets, with repeated training on one half and testing on the other. While the results Dietterich described appear promising, this approach was deemed inappropriate for our data, since training on a data set of 10 cases is certain to lead to low sensitivity and thus would be an inefficient test of any procedure, committee, or single NN.

Our numeric experiments revealed that the committee approach always enjoyed higher sensitivity than any single best network, considered across the several log MSE levels used in training. Equally important, the observed false-positive rate per patient for the committee was 2.9, while it was 4.6 for the average single network; we believe this improvement is clinically meaningful. To summarize, results of our investigation confirm that use of a committee of classifiers such as NNs may bring classification results that are noticeably improved over those achievable with use of a single classifier.

ACKNOWLEDGMENT

We thank C. Daniel Johnson, MD, of the Mayo Clinic for providing CT colonography data sets.

REFERENCES

1. Summers RM, Hara AK, Luboldt W, Johnson CD. CT and MR colonography: summary of progress from 1995 to 2000. Curr Probl Diagn Radiol 2001; 30:147–167.
2. Cherkassky VS, Mulier F. Learning from data: concepts, theory, and methods. New York, NY: Wiley-Interscience, 1998.
3. Summers RM, Beaulieu CF, Pusanik LM, et al. Automated polyp detector for CT colonography: feasibility study. Radiology 2000; 216:284–290.
4. O'Brien MJ. Colorectal polyps. In: O'Brien MJ, Winawer SJ, Waye JD, eds. Gastrointestinal cancer. New York, NY: Gower Medical, 1992; 3.1–3.41.
5. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics. Heidelberg, Germany: Springer-Verlag, 2001.
6. Dietterich TG. Machine learning research: four current directions. AI Magazine 1999; 18:97–136.
7. Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intell Syst 1998; 13:44–49.
8. Nguyen D, Widrow B. Improving the learning speed of two-layer neural networks by choosing initial values of the adaptive weights. In: Proceedings of the Third International Joint Conference on Neural Networks, San Diego, Calif, 1990; 21–26.
9. More JJ. The Levenberg-Marquardt algorithm: implementation and theory. In: Watson GA, ed. Numerical analysis, lecture notes in mathematics 630. Heidelberg, Germany: Springer-Verlag, 1977; 105–116.
10. Tibshirani R, Efron B. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997; 92:548–560.
11. Summers RM, Jerebko AK, Franaszek M, et al. Complementary role of computer-aided detection of colonic polyps with CT colonography. Radiology 2002; 225:391–399.
12. Dietterich T. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998; 10:1895–1923.